Please cite this article as:

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2015). Comparing the percentage of non-overlapping data approach and the hierarchical linear modeling approach for synthesizing single-case studies in autism research. *Research in Autism Spectrum Disorders, 11,* 112-125. doi:10.1016/j.rasd.2014.12.002

Comparing the percentage of non-overlapping data approach and the hierarchical linear modeling approach for synthesizing single-case studies in autism research

Mieke Heyvaert[1, 2], Lore Saenen[1], Bea Maes[1], & Patrick Onghena[1]

[1] Faculty of Psychology and Educational Sciences - KU Leuven

[2] Postdoctoral Fellow of the Research Foundation - Flanders (Belgium)

Correspondence concerning this article can be addressed to Dr. Mieke Heyvaert, Methodology of Educational Sciences Research Group, Tiensestraat 102 - Box 3762, B-3000 Leuven, Belgium. Phone +32 16 326265. E-mail Mieke.Heyvaert@ppw.kuleuven.be

Abstract

We examined the performance of two approaches for synthesizing single-case experimental data: the percentage of non-overlapping data (PND) approach and the hierarchical linear modeling (HLM) approach. The comparison was performed by analyzing an empirical dataset on behavioral interventions for reducing challenging behavior in persons with autism by means of the two approaches. We compared the findings of both approaches for analyzing the outcomes of the behavioral interventions as well as for identifying moderating variables. With respect to the analysis of the interventions' outcomes, similar positive results were found based on both approaches. With respect to the moderating variables, *Functional analysis/assessment* and *Availability of follow up data* were found to be statistically significant moderators by means of the PND as well as the HLM approach. The variables *Intervention type*, *Availability of generalization attempts*, *Design type*, and *Availability of inter-rater reliability data* were also found to be statistically significant moderators by means of the PND approach. The PND approach seems overly liberal in identifying statistically significant predictors in comparison to the HLM approach.

Keywords: single-case research, single-subject experimental designs, meta-analysis, systematic review, behavioral interventions, challenging behavior

Comparing the percentage of non-overlapping data approach and the hierarchical linear modeling approach for synthesizing single-case studies in autism research

1. Introduction

A considerable number of empirical studies on interventions in persons with autism rely on single-case experimental designs (SCEDs) (e.g., Bulkeley, Bundy, Roberts, & Einfeld, 2013; Ganz et al., 2011; Matson, Turygin, Beighley, & Matson, 2012; Reynhout & Carter, 2011; Wang, Cui, & Parrila, 2011). SCEDs are often used to evaluate the effect of an intervention for a single person or a small number of persons, although they can also be used for studying a large number of participants (e.g., Geller, Paterson, & Talbott, 1982). In an SCED involving a single participant, the intervention (e.g., a social stories intervention) can be considered as one of the levels of the independent variable, which is manipulated by the experimenter, and the effect can be evaluated by a dependent variable (e.g., prosocial behavior), which is measured repeatedly for this single person over time.

1.1. Analyzing individual study data

Traditionally, single-case researchers have been using visual analysis for evaluating behavior change, by means of inspecting graphed SCED data for changes in level, variability, trend, latency to change, and overlap between phases in order to judge the reliability and consistency of treatment effects (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Kazdin, 2011). It is concluded that the changes in behavior across phases result from the implemented treatment and are indicative of improvement when the changes in level, trend, and/or variability are in the desired direction and when they are immediate, readily discernible, and maintained over time (Busse, Kratochwill, & Elliott, 1995). However, when there is a long latency between manipulation of the independent variable and change in the dependent variable, when level changes across conditions are small and/or similar to changes within conditions, and when trends do not conform to those predicted following manipulation of the independent variable, demonstration of a functional relationship between the independent and dependent variable is compromised (Horner et al., 2005; Kazdin, 2011).

A group of SCED effect size measures that closely relates to visual analysis are nonoverlap statistics, such as the percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), the percentage of data points exceeding the median of baseline phase (PEM; Ma, 2006), and the percentage of all nonoverlapping data (PAND; Parker,

Hagan-Burke, & Vannest, 2007). These nonoverlap statistics are all nonparametric effect size measures. In addition to nonparametric SCED effect size measures, parametric effect size measures for analyzing and interpreting SCED data have been developed, such as standardized mean difference (SMD) and regression-based effect size measures. Examples are the SMD effect size measure developed by Hedges, Pustejovsky, and Shadish (2012), the piecewise regression approach of Center, Skiba, and Casey (1985-1986), the regression approach of White, Rusch, Kazdin, and Hartmann (1989), the regression approach of Allison and Gorman (1993), and hierarchical linear modeling (HLM; Van den Noortgate & Onghena, 2003a, 2003b).

Next to the use of descriptive statistics (including parametric and nonparametric effect size measures), inferential statistical techniques can be used for analyzing SCED data (including parametric and nonparametric significance tests). Parametric significance tests traditionally used for analyzing group-comparison studies, such as *t*- and *F*-test, are often not appropriate to analyze SCEDs because assumptions of normality are frequently violated for SCED data, SCED data are often autocorrelated, and these tests are insensitive to trends that occur within a phase (Houle, 2009; Smith, 2012). Parametric approaches that are more appropriate to analyze SCED data are for instance generalized least squares regression analysis (Maggin, Swaminathan, et al., 2011), interrupted time series analysis procedures such as ITSACORR (Crosbie, 1993, 1995), piecewise regression analysis (Center et al., 1985-1986), and HLM (Van den Noortgate & Onghena, 2003a, 2003b). Of those parametric approaches, the HLM approach is considered one of the most promising parametric approaches for analyzing SCED data (Gage & Lewis, 2014; Kratochwill et al., 2010; Van den Noortgate & Onghena, 2008; Wolery, Busick, Reichow, & Barton, 2010).

Furthermore, nonparametric significance tests have been recommended for analyzing SCEDs, because they are valid without making distributional assumptions (e.g., Kruskal-Wallis test, Wilcoxon-Mann-Whitney test, randomization test for raw data). An advantage of the randomization test for raw data over the Kruskal-Wallis and Wilcoxon-Mann-Whitney test is that it allows deriving a *p* value without degrading the observed scores to ranks (Onghena & Edgington, 2005). However, the randomization test can only be validly used when the measurement occasions are randomly assigned to the experimental conditions before the start of the experiment, which might not be possible or desirable for all SCEDs (Heyvaert & Onghena, 2014; Onghena & Edgington, 2005).

1.2. Meta-analyzing SCED data

Within the present evidence-based practice movement, researchers, practitioners, and policymakers increasingly rely on research syntheses and meta-analyses to render guidelines for best practice (Beretvas & Chung, 2008; Shadish & Rindskopf, 2007). Important merits of SCED meta-analytic research over individual SCED studies include: a higher statistical power to detect effects, more accurate effect size estimations, the ability to make more convincing generalizations to a larger population, and the ability to identify sources of heterogeneity and to test moderators to explain detected between-study variation. Whereas the analysis of individual SCED studies can be accomplished using visual and/or statistical methods, the synthesis of a large number of SCEDs in a meta-analysis necessitates the use of statistical methods (Smith, 2012).

One frequently used approach for conducting a meta-analysis of SCED studies is to calculate the (weighted) average of the effect sizes of all SCED studies included in the meta-analysis. For instance, nonoverlap effect size measures such as PND, PEM, or PAND are calculated for individual SCED studies and are afterwards aggregated over all SCED studies included in the meta-analysis. Many meta-analyses of SCEDs published in the field of autism research are conducted by aggregating nonoverlap effect sizes. For most of these meta-analyses the PND effect size is used (e.g., Bellini & Akullian, 2007; Campbell, 2003; Preston & Carter, 2009; Tincani & Devis, 2010). In section **1.3** we will discuss in detail how the PND approach can be used for meta-analyzing SCED data.

More advanced approaches for conducting meta-analyses of SCED studies are for instance the Busk and Serlin's (1992) approaches and the HLM approach proposed by Van den Noortgate and Onghena (2003a, 2003b, 2008). In the field of autism research, recently several meta-analyses of SCEDs have been conducted that used the HLM approach (e.g., Vanderkerken, Heyvaert, Maes, & Onghena, 2013; Wang et al., 2011; Wang, Parrila, & Cui, 2013). In section **1.4** we will discuss in detail how the HLM approach can be used for meta-analyzing SCED data.

1.3. Using the PND approach for meta-analyzing SCED data

PND was the first effect size measure proposed for quantitatively synthesizing SCED data. Nowadays, PND is still the most often used effect size index across meta-analyses of SCEDs in the field of disability research (Maggin, O'Keeffe, & Johnson, 2011). PND is a nonparametric effect size measure that aims to calculate the non-overlap between baseline and intervention phases in SCEDs. Single-case researchers using PND have to identify the most

extreme data point in the baseline phase (i.e., the lowest baseline data point if the goal of the SCED is to decrease undesirable behavior, or the highest baseline data point if the goal of the SCED is to increase desirable behavior) and to determine the percentage of intervention phase data points that exceeds this most extreme data point. Accordingly, PND can be calculated by dividing the number of intervention data points that exceeds the most extreme baseline data point in the expected direction by the total number of intervention phase data points (Scruggs et al., 1987). PND effect sizes can range from 0% to 100%. When PND is equal to or larger than 90% the intervention is 'highly effective', when PND is equal to or larger than 70% but smaller than 90% the intervention is 'effective', when PND is equal to or larger than 50% but smaller than 70% the intervention is 'questionable', and when PND is smaller than 50% the intervention is 'ineffective' (Scruggs & Mastropieri, 1998).

First, SCED meta-analysts are often interested in the overall efficacy of an intervention, treatment, or program. When PND is used as an effect size measure in meta-analyses of SCEDs, the PND effect sizes calculated for each individual SCED that is included in the meta-analysis have to be aggregated. Often, the mean of the individual PND effect sizes is calculated as an overall measure of the efficacy of the intervention, treatment, or program. PND effect sizes are often weighted when they are aggregated: (1) When more than one dependent variable is targeted for a participant, the average effect size for that participant is calculated by weighting each dependent variable according to the number of data points reporting on that dependent variable, and (2) within each SCED study, effect sizes are weighted according to the number of data points per participant and then averaged for all participants to yield an effect size per study (Campbell, 2003).

Second, SCED meta-analysts are often interested in variables that moderate the overall efficacy of the intervention, treatment, or program of interest. For the PND approach it is not prescribed which statistical procedure should be used to determine the statistical significance of the predictor variables. Sometimes single-case synthesis authors use parametric statistical tests for studying the significance of predictor variables for the PND approach (e.g., parametric analyses of variance and hierarchical multiple regression analyses; Campbell, 2003). However, more often nonparametric statistical tests are used for studying the significance of predictor variables for the PND approach, such as Kruskal-Wallis analyses of variance (ANOVAs). Using nonparametric statistical tests for the PND approach makes more sense, because PND is a nonparametric effect size measure. Furthermore, it is unlikely that all the assumptions for parametric statistical tests for studying the significance of predictor variables are met when used within the PND approach for synthesizing SCED data.

We refer the reader interested in the details for conducting PND analyses to the work of Scruggs et al. (1987) and Scruggs and Mastropieri (1998, 2013).

1.4. Using the HLM approach for meta-analyzing SCED data

From the parametric approaches for synthesizing SCED data, the HLM approach developed by Van den Noortgate and Onghena (2003a, 2003b, 2008) is considered one of the most promising approaches (Gage & Lewis, 2014; Kratochwill et al., 2010). Important merits of the HLM approach are that it is able to estimate and test mean shift, trend, and variability in SCED data, to calculate robust *t*-ratios using maximum likelihood estimation for significance testing, to account for moderating variables, and to take into account the dependencies that may result from the hierarchical clustering of SCED data (Gage & Lewis, 2014; Kratochwill et al., 2010; Van den Noortgate & Onghena, 2008; Wolery et al., 2010).

The HLM approach is particularly interesting for analyzing SCED data that show a hierarchical structure. An example of a common hierarchical three-level structure is the following: An SCED meta-analytical dataset includes a number of SCED studies, each SCED study reports on one or more participants, and for each participant repeated measurements are reported for the dependent variable of interest. The HLM approach is able to account for the possible dependency that may result from the three-level nesting by modeling the variation within participants, between participants of the same study, and between studies included in the meta-analysis (Van den Noortgate & Onghena, 2008). In other words, the variance in observed intervention effects is split up in sampling variance, variance between participants from the same study, and variance between studies, and the HLM approach tries to explain this variation by the inclusion of case and study characteristics (i.e., the predictors).

Researchers applying the HLM approach for meta-analyzing SCED data can for instance use the SAS procedure MIXED (restricted maximum-likelihood procedure) to estimate and test various parameters of interest, such as the overall intercept, the overall intervention effect, and the covariance parameters (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; Van den Noortgate & Onghena, 2003a, 2003b, 2008).

First, in order to analyze the overall efficacy of an intervention, treatment, or program of interest, the HLM approach can be used to estimate and test the mean intervention effect (i.e., the overall effect). HLM researchers use the Wald test to test the null hypothesis that on average there is no statistically significant effect of the intervention, treatment, or program of interest on the dependent variable of interest. In addition, HLM researchers estimate and test

the variance of the intervention effect between studies and between participants using the restricted maximum-likelihood procedure including the likelihood ratio test.

Second, in order to examine which variables moderate the overall efficacy of the intervention, treatment, or program of interest, predictor variables can be included in the HLM. This extended model can be used to test whether the overall treatment effect depends on the coded predictor variables.

We refer the reader interested in the details and code for conducting HLM analyses to the work of Van den Noortgate and Onghena (2003a, 2003b, 2008) and Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014).

1.5. Objectives of the present study

The PND and the HLM approach are often used for synthesizing SCED data in the field of autism intervention research (cf. **1.2**). However, what is missing for applied researchers, teachers, and students who want to embark on a meta-analysis of SCED studies journey is an empirical comparison of the PND and the HLM approach for conducting meta-analyses of SCEDs. Such a comparison can expose the merits and drawbacks of the PND and the HLM approach, which can help applied researchers, teachers, and students to decide whether they will use the PND or the HLM approach for conducting their meta-analysis of SCEDs.

The present paper can serve another goal as well. During the last decades the number of published meta-analyses of SCED studies is steadily increasing. Autism practitioners and policymakers who want to decide whether or not to start using (or continue using) a certain intervention, treatment, or program can consult published meta-analyses of SCED studies on this intervention, treatment, or program to assess its effects. When making this assessment, it is important to understand the methodology behind the approach used for these meta-analyses, and the assumptions related to the applied method(s). In the present paper we will describe and discuss the methodologies and assumptions behind the PND and the HLM approach. Furthermore, we will discuss the merits and drawbacks of both approaches. The latter will help autism practitioners and policymakers to judge the strengths and the weaknesses of the published SCED meta-analyses they come across.

The first aim of the present study was to compare the performance of the PND and the HLM approach for synthesizing SCED data. We did that by (1) analyzing the same dataset by means of the PND and the HLM approach and comparing the findings, and (2) examining how PND and HLM effect sizes correlate with one another. We used an empirical dataset on

behavioral interventions for reducing challenging behavior in persons with autism (Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014a) for the methodological comparison of the PND and the HLM approach. We compared the findings of both approaches for analyzing the outcomes of the behavioral interventions as well as for identifying moderating variables. The second aim of the present study was to provide applied single-case researchers, practitioners, and policymakers an overview of the merits and drawbacks of the PND and the HLM approach for synthesizing SCED data.

2. Methods

2.1. Dataset

The empirical dataset that was used for the methodological comparison of the PND and the HLM approach was the dataset of Heyvaert et al. (2014a) on behavioral interventions for reducing challenging behavior in people with autism. The following SCED studies and participants were included in this dataset: (1) Included participants were diagnosed with autism, (2) the behavioral intervention described in the SCED study targeted reduction of self-injurious, stereotyped, or disruptive behavior, aggression, or property destruction, (3) only SCEDs were included that described for each participant the level of challenging behavior under baseline and intervention conditions (each condition containing at least two data points) in graphical or tabular format, (4) because the Heyvaert et al. (2014a) dataset was an update of the Campbell (2003) dataset, only SCED studies published between 1999 and 2012 were included, and (5) articles had to be written in English in order to be understood by the research team. The SCED studies were retrieved by systematically searching seven electronic databases, 26 relevant journals, bibliographies of relevant articles, and three citation indexes (see Heyvaert et al., 2014a, for all details). Two hundred and thirteen studies representing 358 persons with autism met the eligibility criteria and were included.

UnGraph Version 5 (Biosoft, 1997-2014) was used to extract raw data (i.e., XY-coordinates) on the behavioral outcomes from the included graphs. The Heyvaert et al. (2014a) dataset was an update of the Campbell (2003) dataset. Campbell (2003) calculated all effect sizes by comparing the first baseline phase to the final treatment phase. In order for the Heyvaert et al. (2014a) dataset to be comparable to the Campbell (2003) dataset, all effect sizes were calculated by comparing the first baseline phase to the final treatment phase. Accordingly, the calculated effect sizes related to behavior change between baseline and treatment phases, and for instance not to generalization or maintenance effects.

Heyvaert et al. (2014a) extracted three groups of variables that were likely to moderate the behavioral interventions' outcomes. The variables were selected based on previous research (Campbell, 2003; Didden, Duker, & Korzilius, 1997; LaGrow & Repp, 1984; Matson, Benavidez, Stabinsky Compton, Paclawskyj, & Baglio, 1996). A first group of coded variables were the participant characteristics *Age, Gender, Criteria used for diagnosing autism, Level of intellectual disability,* and *Level of verbal communication ability*. A second group of coded variables were the intervention characteristics *Intervention type, Targeted challenging behavior, Parental involvement in the intervention, Functional analysis/assessment, Availability of follow up data,* and *Availability of generalization attempts*. A third group of coded variables were the experimental characteristics *Design type, Publication year,* and *Availability of inter-rater reliability data.* Table 1 outlines the codes for the participant, intervention, and experimental variables. Inter-rater agreement between the first and second author on coding the 14 predictor variables was 99.82% (see Heyvaert et al., 2014a, for all details).

INSERT TABLE 1 ABOUT HERE

2.2. Using the PND approach for analyzing the dataset

We first analyzed the outcomes of SCED studies included in the empirical dataset. As a first step, we calculated the PND effect size for each graph that met the inclusion criteria (cf. **2.1**). Next, in order to synthesize the reported outcomes using the PND approach, we aggregated all the calculated PND effect sizes. We weighted the PND effect sizes during the aggregation process: (1) When more than one dependent variable was targeted for a participant, the average effect size for that participant was calculated by weighting each dependent variable according to the number of data points reporting on that dependent variable, and (2) within each SCED study, effect sizes were weighted according to the number of data points per participant and then averaged for all participants to yield an effect size per study (Campbell, 2003).

Second, we studied which variables moderated these treatment outcomes. We used Kruskal-Wallis ANOVAs to examine the coded participant, intervention, and experimental characteristics (cf. **2.1**). All statistical analyses related to the predictor variables were conducted at the participant level, because the coded characteristics related to the participants, and not to the SCED studies. We used SPSS software (Version 22; SPSS Inc., 2013-2014) to conduct the moderator analyses.

2.3. Using the HLM approach for analyzing the dataset

The empirical dataset we used for our methodological comparison showed a hierarchical three-level structure: The dataset included 213 SCED studies, that described 358 individuals with autism and challenging behavior, and for each individual repeated measurements of challenging behavior were reported. Accordingly, using the HLM approach for analyzing this dataset seemed to be a good fit (cf. **1.4**).

First, in order to analyze the outcomes of the behavioral interventions using the HLM approach we estimated and tested the mean intervention effect (i.e., the overall effect). When the mean intervention effect estimated by the HLM approach for studies on interventions aimed at reducing challenging behavior is a negative value, this implies that the level of challenging behavior is on average lower in the intervention conditions, compared to the baseline conditions, so that the intervention works in reducing the challenging behavior. The larger this negative value is for the estimated mean intervention effect, the lower the level of challenging behavior is in the intervention conditions, in comparison with the baseline conditions. Following the HLM approach, we used the Wald test to test the null hypothesis that on average there was no statistically significant effect of the behavioral interventions on the level of challenging behavior. In addition, we estimated and tested the variance of the overall effect between studies and between participants using the restricted maximum-likelihood procedure including the likelihood ratio test.

Second, we examined which variables moderated the outcomes using the HLM approach. When the likelihood ratio test indicates that there is statistically significant between-participant and/or between-study variance of the overall effect, the presence of moderators is likely, and predictor variables can be included in the analyses. Using the HLM approach we tested whether the overall treatment effect depended on the coded participant, intervention, and experimental characteristics (cf. **2.1**). We used SAS 9.3 Software (SAS Institute Inc., 2011-2014) to conduct all HLM analyses.

2.4. Correlations between the PND and the HLM effect sizes

In order to examine the relationship between the PND and the HLM effect sizes, we constructed a scatter plot of the PND and the HLM effect sizes, and we calculated the Pearson’s product-moment and the Kendall’s Tau-b correlation coefficient. All correlational

analyses were conducted at the participant level, using SPSS (Version 22; SPSS Inc., 2013-2014).

3. Results

3.1. Comparing the PND and the HLM approach for examining SCED study outcomes

In order to synthesize the outcomes reported in the SCED studies using the PND approach, the PND effect size was calculated for all included participants and studies. The aggregated PND effect size calculated over all included SCED studies was 75.9%. Following the interpretational guidelines of Scruggs et al. (1987), we concluded for the PND approach that the behavioral interventions were on average effective in reducing challenging behavior.

In order to synthesize the outcomes reported in the SCED studies using the HLM approach, we looked at the three-level random effects regression model without moderators, presented in Table 2. We concluded that the behavioral interventions were on average highly effective: In comparison with the baseline conditions, the level of challenging behavior was 3.92 standard deviations lower in the behavioral intervention conditions. According to the Wald test, this reduction in challenging behavior was statistically significant, $Z = -12.14$, $p < .0001$. However, considering the covariance parameter estimates, the intervention effects showed to vary significantly over the participants (estimated variance of 28.81; $SE = 2.81$; $Z = 10.26$, $p < .0001$). The variance between studies was much smaller than the variance between participants, but also statistically significant (estimated variance of 3.94; $SE = 2.04$; $Z = 1.93$, $p = .0270$). We will examine which moderator variables can explain this variation of intervention effects in the next section.

INSERT TABLE 2 ABOUT HERE

3.2. Comparing the PND and the HLM approach for examining moderating variables

In order to identify moderating variables using the PND approach, we used Kruskal-Wallis ANOVAs to examine the coded participant, intervention, and experimental characteristics. Each moderator was tested separately by means of Kruskal-Wallis ANOVAs. The results of the Kruskal-Wallis ANOVAs for the PND approach are presented in Table 3.

The analyses revealed that there was statistically significant evidence for moderator effects of the characteristics *Intervention type*, $\chi^2 = 35.66$, df = 12, $p < .0001$, *Functional analysis/assessment*, $\chi^2 = 12.90$, df = 1, $p < .0001$, *Availability of follow up data*, $\chi^2 = 3.98$, df = 1, $p = .046$, *Availability of generalization attempts*, $\chi^2 = 9.10$, df = 1, $p = .003$, *Design type*,

χ^2 = 16.59, df = 8, p = .035, and *Availability of inter-rater reliability data*, χ^2 = 6.90, df = 1, p = .009. For the moderator *Intervention type*, the highest effect sizes were found for the *Antecedent exercise only, Differential reinforcement of other behavior only, Combinations of positive interventions, Differential reinforcement of incompatible behavior only, Combinations of positive and aversive interventions, Noncontingent reinforcement only, Social stories only, Mindfulness-based strategy only, Punishment only, Differential reinforcement of alternative behavior only, Antecedent control only, Picture exchange communication system only,* and *Escape only* interventions respectively. For the moderator *Functional analysis/assessment*, higher intervention effects were found for participants for whom a functional analysis or assessment was reported. For the moderator *Availability of follow up data,* higher intervention effects were found for participants for whom follow up data were reported. On the contrary, for the moderator *Availability of generalization attempts*, higher intervention effects were found for participants for whom no generalization attempts were reported. For the moderator *Design type*, the highest effect sizes were found for the categories *Combination of multiple baseline and alternating treatments design*, *Combination of alternating treatments and reversal design*, *Alternating treatments only design*, *Reversal only design*, *Multiple baseline only design*, *Simple A-B only design*, *Combination of A-B and alternating treatments design,* and *Combination of multiple baseline and reversal design* respectively. For the moderator *Availability of inter-rater reliability data*, higher intervention effects were found for participants for whom inter-rater reliability data were reported. Using the PND approach, no statistical evidence was found for moderator effects of the other variables (see Table 3).

INSERT TABLE 3 ABOUT HERE

In order to identify moderating variables using the HLM approach, we tested whether the overall treatment effect depended on the coded participant, intervention, and experimental characteristics. Parallel to the Kruskal-Wallis ANOVAs for the PND approach, we separately added each predictor to the random effects regression model to test its moderating effects for the HLM approach. The results of the moderator analyses for the HLM approach are presented in Table 3.

The analyses revealed that there was statistically significant evidence for moderator effects of the characteristics *Functional analysis/assessment*, Z = 2.45, p = .014, and *Availability of follow up data*, Z = 2.03, p = .043. Parallel to the results for the PND approach,

for the moderator *Functional analysis/assessment* higher intervention effects were found for participants for whom a functional analysis or assessment was reported, and for the moderator *Availability of follow up data* higher intervention effects were found for participants for whom follow up data were reported. Using the HLM approach, no statistical evidence was found for moderator effects of the other variables (see Table 3).

3.3. Correlations between the PND and the HLM effect sizes

In order to examine the correlation between the PND and the HLM effect sizes, we first plotted the relationship between the PND and the HLM effect sizes calculated at the participant level. As we mentioned before, when the intervention effect estimated by the HLM approach for studies on interventions aimed at reducing challenging behavior is a negative value, this implies that the intervention works in reducing the challenging behavior. The greater this negative value is for the estimated intervention effect, the lower the level of challenging behavior is in the intervention conditions, in comparison with the baseline conditions. However, for plotting the relationship between the PND and the HLM effect sizes calculated at the participant level, we thought that it would make more sense to the reader if we would invert the calculated HLM effect sizes. By inverting the calculated HLM effect sizes, we simplify the interpretation of the scatter plot: For the PND as well as the HLM approach a larger (positive) effect size corresponds to an intervention that is more effective in reducing the challenging behavior. As can be seen in Figure 1, the PND effect sizes range between 0% and 100%, and the (inversed) observed outcomes for HLM range between -33 and 171.56. For PND, Figure 1 shows clear ceiling effects (i.e., a PND effect size of 100%), but also floor effects (i.e., a PND effect size of 0%). Based on Figure 1 we can preliminary conclude that there was a positive, moderately strong correlation between the PND and the inverted HLM effect sizes.

Next, we used statistical parametric and nonparametric correlation analyses to examine the relationship between the PND and the HLM effect sizes at the participant level. The Pearson's product-moment correlation between the PND and the HLM effect sizes was positive and statistically significant, $r = .389$, $p < .0001$. The Kendall's Tau-b correlation between the PND and the HLM effect sizes was also positive and statistically significant, $\tau = .289$, $p < .0001$. Accordingly, we could for both the Pearson's product-moment and the Kendall's Tau-b correlation analyses reject the null hypothesis that there was no relationship between the PND and the HLM effect sizes. Based on the Pearson's product-moment and the Kendall's Tau-b correlation coefficients, we concluded that the relationship between the PND

and the HLM effect sizes calculated at the participant level was positive, but only moderately strong.

INSERT FIGURE 1 ABOUT HERE

4. Discussion

This study aimed to compare the PND and the HLM approach for synthesizing SCED data. Both approaches are frequently used in published meta-analyses of SCEDs conducted in the field of intervention research in people with autism (e.g., Reynhout & Carter, 2011; Wang et al., 2011). We used an empirical dataset on behavioral interventions for reducing challenging behavior in persons with autism (Heyvaert et al., 2014a) to make the methodological comparison. We compared the findings of both approaches for analyzing the outcomes of the behavioral interventions as well as for identifying moderating variables. In this final section, we will first discuss the findings of the PND and the HLM approach and review the merits and drawbacks of both approaches for synthesizing SCED data. Second, we will discuss the limitations of the present study. Third, we will discuss our study's implications for research, policy, and practice.

4.1. Merits and drawbacks of the PND and the HLM approach for synthesizing SCED data and discussion of the results

Important advantages of the PND approach for synthesizing SCED data are that the PND effect size measure is a very intuitive index that strongly relates to visual analysis results, and that it is easy to interpret. Another important advantage is that the PND index can easily be calculated manually for uncrowded SCED graphs and graphs with minimal overlap between baseline and intervention data points (in that case, only a ruler can be used). However, PND calculation can be complicated when assessing more compacted SCED graphs with more overlap between baseline and intervention data points (Parker & Vannest, 2009). In that case the use of data extraction software might be imperative. The HLM approach requests for *all* SCED graphs that *all* raw data points are extracted. This can be accomplished by using data extraction software. In addition to commercial software that has been developed for extracting raw data from graphs (e.g., UnGraph, developed by Biosoft, 1997-2014), there also exists free software for extracting raw data from graphs (e.g., Bulté & Onghena, 2012). Applied researchers intending to meta-analyze SCED data can consider the additional data

extraction step for *all* SCED graphs to be an important disadvantage of the HLM approach, compared to the PND approach.

Major drawbacks of the PND approach are that (1) the PND index only takes into account one data point from the baseline phase, being the most extreme data point, which implies that PND results can easily be distorted by the presence of outliers in SCED datasets; (2) the PND approach is not able to detect changes in trend in SCED data, and is not able to take into account trend observed in the baseline phase; and (3) the PND index lacks a known sampling distribution, making it impossible to construct a confidence interval or to derive a *p* value using a parametric significance test (Parker & Vannest, 2009). Because of these drawbacks, and based on empirical and simulation studies that showed that PND has unacceptably high levels of errors, single-case researchers are often dissuaded from using PND for conducting SCED meta-analyses (e.g., Kratochwill et al., 2010; Wolery et al., 2010). However, Gage and Lewis (2014) recently found very similar results for PND and parametric effect size measures: Although the calculation of PND for each individual graph revealed a number of issues, including a bias when one single data point was an outlier in the baseline phase, they found that the PND and the parametric effect size results were congruent at an aggregate level. Because of its ease of use and interpretation, PND is still the most widely used effect size measure for conducting SCED meta-analyses in the field of disability research (Maggin, O'Keeffe, et al., 2011; Scruggs & Mastropieri, 2013).

In contrast to the PND approach, advantages of the HLM approach are that it (1) takes into account all baseline data points and is less easily distorted by the presence of outliers in SCED datasets in comparison with PND; (2) is able to deal with trends, by using an extension to the basic HLM approach (Van den Noortgate & Onghena, 2003b; Moeyaert, Ferron, et al., 2014); and (3) generates not only estimates of the overall treatment effect, the average baseline level, and between-case (co)variance and between-study (co)variance of the intervention effect, but also includes statistical significance tests for these parameters. An additional merit of the HLM approach is that it is able to account for the possible dependency that may result from three-level nesting by modeling the variation within participants, between participants of the same study, and between studies included in the meta-analysis (Van den Noortgate & Onghena, 2008).

The major drawback of the HLM approach is that it is not nearly as intuitive as the PND approach in its calculation and interpretation, as will be discussed further below. The basic HLM approach might already be technically challenging (Moeyaert, Ferron, et al., 2014). When the SCED data require adjustments to the basic HLM procedure, for instance in

order to take into account external events (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013), to take into account different SCED types (Moeyaert, Ugille, et al., in press), or to take into account trend in SCED data (Van den Noortgate & Onghena, 2003b), HLM analyses might become even more challenging in their calculation and interpretation.

When synthesizing the outcomes reported in the SCED studies using the PND approach, an important drawback of the PND approach is that it only provides a descriptive measure of an overall effect. The PND approach does not imply a statistical test that can be used to determine the statistical significance of the overall intervention effect, nor a procedure for calculating a 95% confidence interval. The HLM approach developed by Van den Noortgate and Onghena (2003a, 2003b, 2008) on the contrary offers a lot of descriptive as well as inferential information. The HLM approach can be used to estimate and test the average baseline level (i.e., the overall intercept) and the average intervention effect (i.e., the overall effect) across the included cases and studies. The average baseline level can be used as an indicator for the need for the intervention: If a single-case researcher for instance intends to conduct a social story intervention to increase the level of prosocial behavior of a participant, but this level of prosocial behavior is already high under the baseline condition, it might not be needed to intervene. The average intervention effect indicates the estimated magnitude of the shift in the dependent variable (i.e., challenging behavior) that tends to occur with the intervention. In addition to estimating the average baseline level and the average intervention effect, the HLM approach can be used to estimate measures of between-case (co)variance and between-study (co)variance. From the (co)variance estimates, especially the between-case and between-study variance in the treatment effect can be relevant to the single-case researcher: This measure can be used to determine whether the shift in the dependent behavior associated with the intervention is similar across the included participants and/or studies or whether the intervention is differentially effective over the participants and/or studies (Moeyaert, Ferron, et al., 2014). Nevertheless, single-case synthesis authors using the HLM approach are usually primarily interested in the average treatment effect over the included cases and studies (i.e., the overall effect). Regarding the overall effect, the HLM approach uses the Wald test to test the null hypothesis that on average there is no statistically significant effect of the independent variable on the dependent variable. Regarding the (co)variance estimates, the HLM approach uses the likelihood ratio test to test the (co)variance of the intervention effect at the between-participant and the between-study level. The HLM assumptions we made for analyzing our empirical dataset were that there were no time trends and that the residuals at the three levels were independent, identically, and normally distributed. However, using

extensions of the basic HLM approach, it is possible to for instance model time trends in the intervention phase, and to model dependence between the residuals at the first level (i.e., autocorrelation) (see Moeyaert, Ferron, et al., 2014).

Compared to the HLM approach, an advantage of the PND approach is that Scruggs et al. (1987) provided clear interpretational guidelines to determine whether the interventions under study are 'highly effective', 'effective', 'questionable', or 'ineffective' based on the descriptive effect size index (see section **2.2**). For HLM such interpretational guidelines for the descriptive effect size index are lacking. The interpretation of the overall effect is not so intuitive for the HLM approach: The HLM overall effect size refers to an estimation of the difference between baseline and treatment means, and is presented in standard deviation units. When this mean intervention effect estimated by the HLM approach for studies on interventions aimed at reducing undesirable behavior is a negative value, this implies that the level of undesirable behavior is on average lower in the intervention conditions, compared to the baseline conditions, so that the intervention works in reducing the undesirable behavior. The greater this negative value is for the estimated mean intervention effect, the lower the level of undesirable behavior is in the intervention conditions, in comparison with the baseline conditions.

Looking at the results for the analyses of the interventions' outcomes (cf. **3.1**), the overall PND effect size was 75.9%. Following the interpretational guidelines of Scruggs et al. (1987), we concluded for PND that the behavioral interventions were on average effective in reducing challenging behavior. The average intervention effect estimated by the HLM approach was -3.92, indicating that in comparison to the baseline conditions, the level of challenging behavior was 3.92 standard deviations lower in the behavioral intervention conditions. According to the Wald test included in the HLM approach, this reduction in challenging behavior was statistically significant. These PND and HLM findings correspond to the findings of other meta-analyses published in the domain of persons with disabilities: Behavioral interventions are on average effective in reducing challenging behavior in individuals with disabilities (Didden, Korzilius, van Oorsouw, & Sturmey, 2006; Harvey, Boer, Meyer, & Evans, 2009; Heyvaert, Maes, & Onghena, 2010; Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012; Heyvaert, Saenen, Maes, & Onghena, 2014b; Vanderkerken et al., 2013). Nonetheless, based on the likelihood ratio test included in the HLM approach, we found that the intervention effects varied significantly over the participants and studies included in the dataset. Next, we studied which moderator variables could explain this variation of outcomes.

With respect to the moderating variables analyses, we already mentioned that for the PND approach it is not prescribed which statistical procedure should be used to determine the statistical significance of predictor variables. Although some single-case synthesis authors use parametric statistical tests for studying the significance of predictor variables for the PND approach (e.g., parametric analyses of variance and hierarchical multiple regression analyses; Campbell, 2003), we believe that it is more appropriate to use nonparametric statistical tests for this purpose, such as Kruskal-Wallis ANOVAs. Our rationale is twofold: (1) using nonparametric statistical tests for the PND approach makes more sense because PND is a nonparametric effect size measure, and (2) it is unlikely that all the assumptions for parametric statistical tests for studying the significance of predictor variables are met when used within the PND approach for synthesizing SCED data.

The HLM approach developed by Van den Noortgate and Onghena (2003a, 2003b, 2008) allowed including and testing predictors that explain the between-case and between-study variance of the intervention effect, using the SAS procedure MIXED. Accordingly, we used the HLM approach to test whether the overall treatment effect depended on the coded participant, intervention, and experimental characteristics. Parallel to the Kruskal-Wallis ANOVAs for the PND approach, we separately added each predictor to the random effects regression model to test its moderating effects for the HLM approach.

Looking at the results for the moderating variables analyses (cf. **3.2**), the variables *Functional analysis/assessment* and *Availability of follow up data* were found to be statistically significant moderators of the overall intervention effect by means of the PND as well as the HLM approach. The variable *Functional analysis/assessment* provides an indication of the degree of quality of the behavioral intervention: Planning and conducting a solid functional analysis or assessment provides information on the function or the meaning of the challenging behavior, which can be taken into account when developing and conducting the intervention. This finding corresponds to the findings of the SCED meta-analyses of Campbell (2003), Didden et al. (2006), and Harvey et al. (2009).

For the variable *Availability of follow up data* it is not immediately clear why an SCED study that includes follow up data would result in a greater reduction of challenging behavior during the time the intervention is administered than an SCED study that does not include follow up data. We think of four possible hypotheses why the variable *Availability of follow up data* could moderate the overall intervention effect. First, this effect might be due to an overall planning effect. The follow up measurements are planned before the intervention is implemented and introduced to the participant. A good overall planning (e.g., including

functional analysis/assessment and follow up measurements) might result in a larger positive intervention effect. Second, this moderating effect might be due to an overall reporting effect. Careful reporting of an SCED study might go hand in hand with a more careful approach of the SCED study, and a more careful approach might go hand in hand with better care for the participant and a higher effectiveness of the intervention. Third, in case the SCED researcher expects larger immediate treatment effects, he might find it more worthwhile to plan follow up measurements. Fourth, in case the SCED researcher observes larger immediate treatment effects, he might find it more worthwhile to also report on follow up measurements. Or the other way around: In case the SCED researcher does not find immediate treatment effects (i.e., short term effects), he might not go through the trouble of additionally studying longer term effects (i.e., using follow up measurements). It is also possible that a combination of two or more of these hypotheses would explain why the variable *Availability of follow up data* moderated the overall intervention effect. In any case, more converging evidence is needed before we can convincingly interpret this moderating effect of the availability of follow up data.

Furthermore, the variables *Intervention type*, *Availability of generalization attempts*, *Design type*, and *Availability of inter-rater reliability data* were found to be statistically significant moderators of the overall intervention effect by means of the PND approach, but not by the HLM approach. Accordingly, we conclude that the PND approach including Kruskal-Wallis ANOVAs is overly liberal in identifying statistically significant predictors, in comparison to the HLM approach.

Finally, we examined the correlation between the PND and the HLM effect sizes calculated at the participant level (cf. **3.3**). We generated a scatter plot and used statistical parametric and nonparametric correlation analyses to examine the relationship between the PND and the HLM effect sizes. The Pearson's product-moment as well as the Kendall's Tau-b correlation index between the PND and the HLM effect sizes were positive and statistically significant. However, the relationship between the PND and the HLM effect sizes was only moderately strong. Accordingly, we conclude that the PND and the HLM effect sizes measure related, but not similar, effects.

Because there is not yet a consensus on which statistical method(s) should be preferred for conducting meta-analyses of SCED data, methodologists advise to do sensitivity analyses by reporting on (one or more) nonparametric and parametric approaches for conducting meta-analyses of SCED data and to afterwards compare results over the approaches to see whether they yield consistent results (Kratochwill et al., 2010). Consistency of results can relate to the

effectiveness of the studied interventions, as well as to the question which variables moderate intervention effectiveness. For the empirical dataset on behavioral interventions for reducing challenging behavior in persons with autism of Heyvaert et al. (2014a), there was consistency between the PND and the HLM approach on the effectiveness of the behavioral interventions in reducing challenging behavior. The HLM approach, but not the PND approach, allowed to additionally examine the statistical significance of the overall intervention effect and the variance of the overall intervention effect over the included participants and SCED studies. Based on the HLM approach, we concluded that the reduction in challenging behavior due to the interventions was statistically significant, but that the intervention effects varied significantly over the participants and studies included in the dataset. For the predictor analyses, there was consistency between the PND and the HLM approach for the variables *Functional analysis/assessment* and *Availability of follow up data:* Both were found to be statistically significant moderators of the intervention effect. However, for the variables *Intervention type*, *Availability of generalization attempts*, *Design type*, and *Availability of inter-rater reliability data* there was no consistency between the PND and the HLM approach: These four variables were only found to be statistically significant moderators by the PND approach.

4.2. Limitations

Our study's results must be interpreted in light of its limitations. First of all, the primary objective of the present study was not to determine the overall efficacy of behavioral interventions in reducing challenging behavior in individuals with autism and to determine which variables influence the overall efficacy of these interventions. The present study only had a methodological focus: We intended to compare the PND and the HLM approach as methods for synthesizing SCED data. When it would have been our intent to answer the substantive research questions on the overall efficacy of behavioral interventions in reducing challenging behavior in individuals with autism and on determining which variables influence this overall efficacy, we should have used criteria for evaluating the quality of each SCED study included in the meta-analysis. For instance, Kratochwill et al. (2010) stipulated in their *Standards for SCED data* that an SCED must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions. Examples of SCEDs meeting this standard are for instance ABAB designs and multiple baseline designs with at least three baseline conditions. Examples of SCEDs not meeting this standard are for instance AB, ABA, and BAB designs. Kratochwill et al. (2010) stipulated

that calculation of SCED effect sizes to determine efficacy is relevant only if the included SCED studies are of sufficient quality to be confident in the findings obtained for the individual SCEDs. The empirical dataset we used for the comparison of the PND and the HLM approach did not answer to these standards: First and foremost, this dataset only included data points from the first baseline phase and the final intervention phase from the included graphs (Heyvaert et al., 2014a). This simplification procedure implied that only AB contrasts were included in the dataset. Furthermore, this simplification procedure implied that some contrasts and data points were not included in the dataset. If we would have wanted to answer the substantive research questions relating to the overall efficacy of behavioral interventions in reducing challenging behavior in individuals with autism and to the variables moderating the overall efficacy of these interventions, we should have applied inclusion criteria based on the *Standards* of Kratochwill et al. (2010).

Second, we only included the PND and the HLM approach in our comparison of frequently used approaches for synthesizing SCED studies. Several other nonparametric and parametric approaches for synthesizing SCED studies have recently been developed, and it might be interesting to include them in the comparison too. For instance, several new nonparametric effect size measures have been developed recently, such as Tau_{novlap} and Tau-*U* (Parker, Vannest, & Davis, 2011), but an empirical comparison of their statistical properties to other SCED effect size measures is lacking for the time being. Furthermore, Hedges et al. (2012) recently developed an SMD effect size measure for SCEDs that is said to be directly comparable with Cohen's *d*. Future research might focus on the comparison between the HLM approach developed by Van den Noortgate and Onghena (2003a, 2003b, 2008) and Hedges et al. (2012)'s recently developed SMD effect size, and study what the differences are between both approaches, and whether one approach should be preferred over the other (under which circumstances).

4.3. Implications

Our article can be valuable for several groups of stakeholders in the field of autism spectrum disorders (ASD). First, a considerable number of ASD researchers use SCEDs for evaluating the effects of interventions, treatments, and programs. When ASD researchers want to obtain a higher statistical power to detect effects, produce more accurate effect size estimations, make more convincing generalizations to a larger population, and test moderators to explain between-participant and/or between-study variation, they benefit from synthesizing SCED data in a meta-analysis. Often used methods for synthesizing SCED data in the field of

ASD intervention research are the PND and the HLM approach. However, what was missing for ASD researchers who wanted to conduct a meta-analysis of SCED studies was an empirical comparison of the PND and the HLM approach for conducting meta-analyses of SCEDs. In the present study we compared the PND and the HLM approach by using both approaches to analyze a single empirical dataset, and we exposed the merits and drawbacks of both approaches. Accordingly, our study can help ASD researchers to decide whether they will use the PND or the HLM approach for conducting their SCED meta-analysis.

Second, ASD practitioners and policymakers increasingly rely on meta-analytical evidence to render guidelines for best practice. We provided more information on the methodology and assumptions behind two approaches that are often used in the field of ASD for meta-analyzing empirical SCED studies (i.e., the PND and the HLM approach). We also discussed the merits and drawbacks of the PND and the HLM approach. This information can help ASD practitioners and policymakers to better understand published SCED meta-analyses using the PND and the HLM approach, to better assess the strengths and the weaknesses of these meta-analyses, and to distinguish sound from poor SCED meta-analyses. After all, ASD policymakers need to minimize the risk that flawed and misleading study results are used on a large scale to guide practice, and ASD practitioners want to use the best available empirical evidence to inform their daily practice.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research & Therapy, 31,* 621–631. doi:10.1016/0005-7967(93)90115-B

Bellini, S., & Akullian, J. (2007). A meta-analysis of video modeling and video self-modeling interventions for children and adolescents with autism spectrum disorders. *Exceptional Children, 73*, 264–287.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2,* 129–141. doi:10.1080/17489530802446302

Biosoft (1997-2014). *UnGraph Version 5* [Computer software]. Retrieved from http://www.biosoft.com/w/ungraph.htm

Bulkeley, K., Bundy, A., Roberts, J., & Einfeld, S. (2013). ASD intervention research in real world contexts: Refining single case designs. *Research in Autism Spectrum Disorders, 7,* 1257–1264. doi:10.1016/j.rasd.2013.07.014

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes. A software tool for the visual analysis of single-case experimental data. *Methodology, 8,* 104–114. doi:10.1027/1614-2241/a000042

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33,* 269–285. doi:10.1016/0022-4405(95)00014-D

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behaviour in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24,* 120–138. doi:10.1016/S0891-4222(03)00014-3

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387–400. doi:10.1177/002246698501900404

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61,* 966–974. doi:10.1037/0022-006X.61.6.966

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.

Didden, R., Duker, P. C., & Korzilius, H. (1997). Meta-analytic study on treatment effectiveness for problem behaviors with individuals who have mental retardation. *American Journal on Mental Retardation, 101,* 387–399.

Didden, R., Korzilius, H., van Oorsouw, W., & Sturmey, P. (2006). Behavioural treatment of challenging behaviours in individuals with mild mental retardation: Meta-analysis of single-subject research. *American Journal on Mental Retardation, 111,* 290–298.

Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *Journal of Special Education, 48,* 3–16. doi:0022466912443894

Ganz, J. B., Earles-Vollrath, T. L., Mason, R. A., Rispoli, M. J., Heath, A. K., & Parker, R. I. (2011). An aggregate study of single-case research involving aided AAC: Participant characteristics of individuals with autism spectrum disorders. *Research in Autism Spectrum Disorders, 5,* 1500–1509. doi:10.1016/j.rasd.2011.02.011

Geller, E. S., Paterson, L., & Talbott, E. (1982). A behavioral analysis of incentive prompts for motivating seat belt use. *Journal of Applied Behavior Analysis, 15,* 403–413. doi:10.1901/jaba.1982.15-403

Harvey, S. T., Boer, D., Meyer, L. H., & Evans, I. M. (2009). Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice. *Journal of Intellectual & Developmental Disability, 34,* 67–80. doi:10.1080/13668250802690922

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3,* 224–239. doi:10.1002/jrsm.1052

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomization tests for measures of effect size. *Neuropsychological Rehabilitation, 24,* 507–527. doi:10.1080/09602011.2013.818564

Heyvaert, M., Maes, B., & Onghena, P. (2010). A meta-analysis of intervention effects on challenging behaviour among persons with intellectual disabilities. *Journal of Intellectual Disability Research, 54,* 634–649. doi:10.1111/j.1365-2788.2010.01291.x

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33,* 766–780. doi:10.1016/j.ridd.2011.10.010

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014a). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35,* 2463–2476. doi:10.1016/j.ridd.2014.06.017

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014b). Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Focus on effectiveness in single-case experiments. *Journal of Applied Research in Intellectual Disabilities, 27,* 493–590. doi:10.1111/jar.12094

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165–179.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35,* 269–290. doi:10.1353/etc.2012.0011

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (3rd ed.) (pp. 271–305). Boston, MA: Allyn & Bacon.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

LaGrow, S. J., & Repp, A. C. (1984). Stereotypic responding: A review of intervention research. *American Journal of Mental Deficiency, 88,* 595–609.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS© system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30,* 598–617. doi:10.1177/0145445504272974

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19,* 109–135. doi:10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301–321. doi:10.1016/j.jsp.2011.03.004

Matson, J. L., Benavidez, D. A., Stabinsky Compton, L., Paclawskyj, T., & Baglio, C. (1996). Behavioral treatment of autistic persons: A review of research from 1980 to the present. *Research in Developmental Disabilities, 17,* 433–465. doi:10.1016/S0891-4222(96)00030-3

Matson, J. L., Turygin, N. C., Beighley, J., & Matson, M. L. (2012). Status of single-case research designs for evidence-based practice. *Research in Autism Spectrum Disorders, 6,* 931–938. doi:10.1016/j.rasd.2011.12.008

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology, 52,* 191–211. doi:10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods, 45,* 547–559. doi:10.3758/s13428-012-0274-1.

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., Beretvas, S., & Van den Noortgate, W. (in press). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*. doi:10.1037/spq0000068

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56–68. doi:10.1097/00002508-200501000-00007

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40,* 194–204. doi:10.1177/00224669070400040101

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357–367.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303–322. doi:10.1177/0145445511399147

Preston, D., & Carter, M. (2009). A review of the efficacy of the picture exchange communication system intervention. *Journal of Autism and Developmental Disorders, 39,* 1471-1486. doi:10.1007/s10803-009-0763-y

Reynhout, G., & Carter, M. (2011). Evaluation of the efficacy of Social Stories™ using three single subject metrics. *Research in Autism Spectrum Disorders, 5,* 885–900. doi:10.1016/j.rasd.2010.10.003

SAS Institute Inc. (2011-2014). *SAS 9.3 Software* [Computer software]. Cary, NC: SAS Institute Inc.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. Methodology and validation. *Remedial and Special Education, 8,* 24–33. doi:10.1177/074193258700800206

Scruggs, T. E., & Mastropieri, M. A. (1998). Synthesizing single subject research: Issues and applications. *Behavior Modification, 22,* 221–242. doi:10.1177/01454455980223001

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*, 9–19. doi:10.1177/0741932512440730

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95–109. doi:10.1002/ev.217

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17,* 510–550. doi:10.1037/a0029312

SPSS Inc. (2013-2014). *SPSS for Windows Version 22* [Computer software]. Chicago, IL: SPSS Inc.

Tincani, M., & Devis, K. (2010). Quantitative synthesis and component analysis of single-participant studies on the Picture Exchange Communication System. *Remedial and Special Education, 32,* 458–470. doi:10.1177/0741932510362494

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18,* 325–346. doi:10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effects sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35,* 1–10. doi:10.3758/BF03195492

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-based Communication Assessment and Intervention, 2,* 142–151. doi:10.1080/17489530802505362

Vanderkerken, L., Heyvaert, M., Maes, B., & Onghena, P. (2013). Psychosocial interventions for reducing vocal challenging behaviour in persons with autistic disorder: A multilevel meta-analysis of single-case experiments. *Research in Developmental Disabilities, 34,* 4515–4533. doi:10.1016/j.ridd.2013.09.030

Wang, S. Y., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: A meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5,* 562–569. doi:10.1016/j.rasd.2010.06.023

Wang, S. Y., Parrila, R., & Cui, Y. (2013). Meta-analysis of social skills interventions of single-case research for individuals with autism spectrum disorders: Results from three-level HLM. *Journal of Autism and Developmental Disorders, 43,* 1701–1716. doi:10.1007/s10803-012-1726-2

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment, 11,* 281–296.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44,* 18–28. doi:10.1177/0022466908328009

Table 1
Codes for the Participant, Intervention, and Experimental Variables.

Variables	Codes
Participant characteristics:	
Age	≤5 (*n* = 89)
	6-8 (*n* = 106)
	9-12 (*n* = 82)
	≥13 (*n* = 76)
Gender	Male (*n* = 286)
	Female (*n* = 67)
	Not reported (*n* = 5)
Criteria used for diagnosing autism	DSM-III or DSM-III-TR (*n* = 1)
	DSM-IV or DSM-IV-TR (*n* = 30)
	Participant simply described as "autistic" or "had a diagnosis of autism" (*n* = 324)
	ICD-10 (*n* = 3)
Level of intellectual disability	None (IQ > 70; (*n* = 20)
	Mild (70 – 55; *n* = 14)
	Moderate (54 – 40; *n* = 30)
	Severe / Profound (IQ < 40; *n* = 48)
	Not reported / Unclear (*n* = 246)
Level of verbal communication ability	Average language skills (*n* = 27)
	Minimally verbal (*n* = 115)
	Nonverbal (*n* = 61)
	Not reported / Unclear (*n* = 155)
Intervention characteristics:	
Intervention type	Aversive and positive combinations (*n* = 74)
	Positive combinations (*n* = 114)
	Punishment only (aversive; *n* = 10)
	Differential reinforcement of other behavior only (positive; *n* = 14)
	Antecedent control only (positive; *n* = 61)
	Differential reinforcement of incompatible behavior only (positive; *n* = 5)
	Differential reinforcement of alternative behavior only (positive; *n* = 23)
	Antecedent exercise only (positive; *n* = 2)
	Noncontingent reinforcement only (positive; *n* = 14)
	Escape only (positive; *n* = 1)
	Social stories only (positive; *n* = 29)
	Picture exchange communication system only (positive; *n* = 5)
	Mindfulness-based strategy only (positive; *n* = 6)
Targeted challenging behavior	Internal and external combined (*n* = 78)
	External combined (*n* = 37)
	Internal combined (*n* = 1)
	Stereotyped behavior only (internal; *n* = 100)
	Self-injurious behavior only (internal; *n* = 33)

	Disruptive behavior only (external; *n* = 82)
	Aggression only (external; *n* = 26)
	Property destruction only (external; *n* = 1)
Parental involvement in the intervention	Yes (*n* = 56)
	No / Not reported (*n* = 302)
Functional analysis/ assessment	Yes (*n* = 257)
	No / Not reported (*n* = 101)
Availability of follow up data	Yes (*n* = 74)
	No / Not reported (*n* = 284)
Availability of generalization attempts	Yes (*n* = 133)
	No / Not reported (*n* = 225)
Experimental characteristics:	
Design type	Reversal only (*n* = 128)
	Multiple baseline only (*n* = 112)
	Simple A-B only (*n* = 16)
	Multiple baseline + Reversal (*n* = 8)
	Alternating treatments only (*n* = 33)
	Multiple baseline + Alternating treatments (*n* = 7)
	Alternating treatments + Reversal (*n* = 34)
	Simple A-B + Alternating treatments (*n* = 19)
	Multiple baseline + Reversal + Alternating treatments (*n* = 1)
Publication year	1999-2005 (*n* = 114)
	2006-2012 (*n* = 244)
Availability of inter-rater reliability data	Yes (*n* = 333)
	No / Not reported (*n* = 25)

Table 2
Parameter Estimates and Standard Errors for the Raw-data HLM Analyses.

Parameters	Parameter Estimates (Standard Errors)
Fixed effects	
Mean intervention effect	-3.92 (0.32)**
Variance of intervention effect	
Between studies	3.94 (2.04)*
Between participants	28.81 (2.81)**

** p < .05; ** = p < .001*

Table 3

Moderator Analyses for the PND and the HLM Approach.

Variable	PND approach			HLM approach	
	χ^2	DF	*p* value	*Z*	*p* value
Age	2.30	3	.512	-1.80	.072
Gender	0.42	2	.812	0.19	.851
Criteria used for diagnosing autism	3.00	3	.392	-0.73	.464
Level of intellectual disability	9.43	4	.051	0.26	.794
Level of verbal communication ability	1.28	3	.734	-0.55	.582
Intervention type	35.66	12	<.0001***	1.27	.205
Targeted challenging behavior	2.77	7	.905	-0.37	.709
Parental involvement in the intervention	1.52	1	.218	-0.95	.341
Functional analysis/assessment	12.90	1	<.0001***	2.45	.014*
Availability of follow up data	3.98	1	.046*	2.03	.043*
Availability of generalization attempts	9.10	1	.003**	-1.57	.116
Design type	16.59	8	.035*	1.26	.208
Publication year	0.004	1	.952	0.44	.658
Availability of inter-rater reliability data	6.90	1	.009**	0.65	.517

** p < .05; ** = p < .01; *** = p < .001*

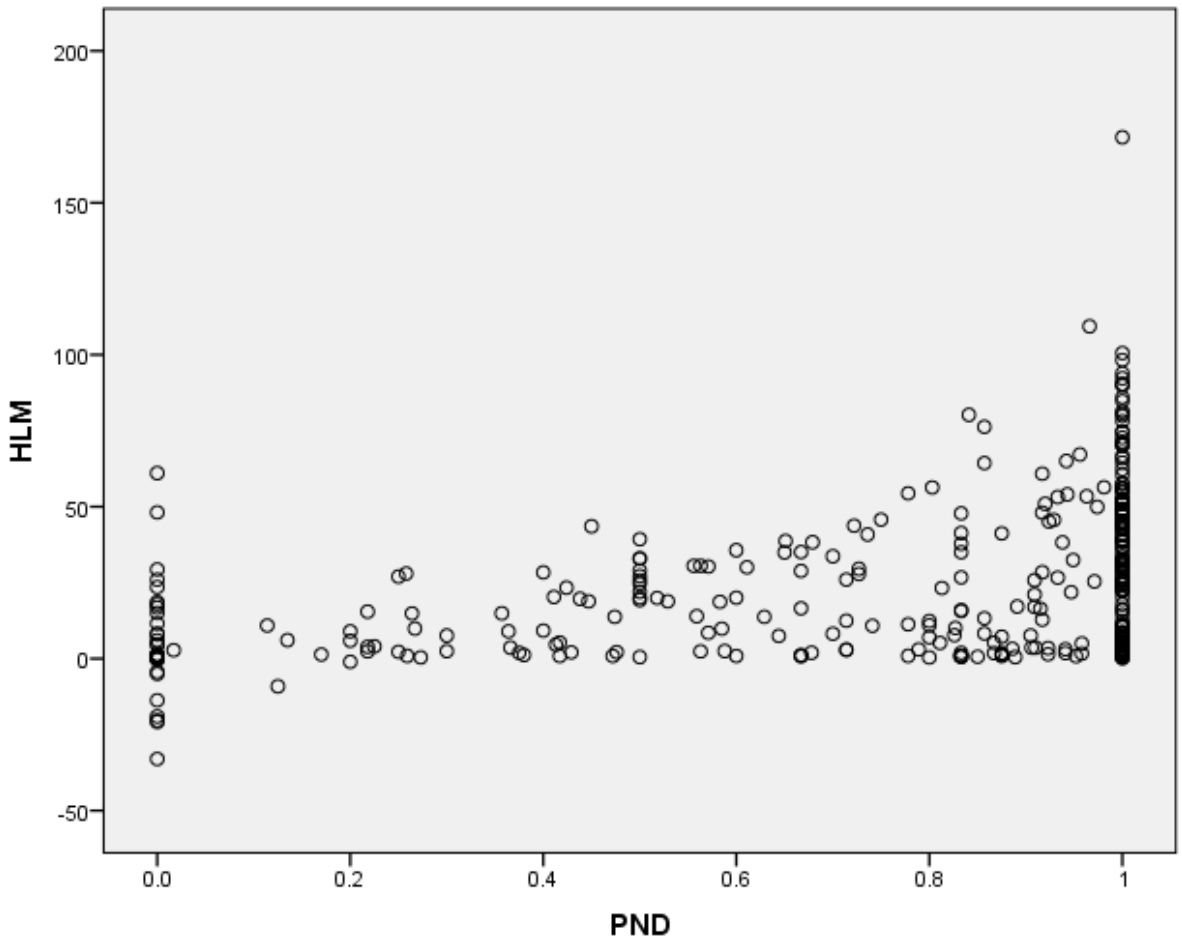

Figure 1. Scatter plot between the PND and the HLM effect sizes at the participant level.